Genetic Epidemiology 28: 138–156 (2005)

Exploiting Gene-Environment Independence in Family-Based Case-Control Studies: Increased Power for Detecting Associations, Interactions and Joint Effects

Nilanjan Chatterjee,[1]* Zeynep Kalaylioglu,[2] and Raymond J. Carroll[3]

[1]*Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Rockville, Maryland*
[2]*Information Management System, Rockville, Maryland*
[3]*Texas A&M University, College Station, Texas*

Family-based case-control studies are popularly used to study the effect of genes and gene-environment interactions in the etiology of rare complex diseases. We consider methods for the analysis of such studies under the assumption that genetic susceptibility (*G*) and environmental exposures (*E*) are independently distributed of each other within families in the source population. Conditional logistic regression, the traditional method of analysis of the data, fails to exploit the independence assumption and hence can be inefficient. Alternatively, one can estimate the multiplicative interaction between *G* and *E* more efficiently using cases only, but the required population-based *G-E* independence assumption is very stringent. In this article, we propose a novel conditional likelihood framework for exploiting the within-family *G-E* independence assumption. This approach leads to a simple and yet highly efficient method of estimating interaction and various other risk parameters of scientific interest. Moreover, we show that the same paradigm also leads to a number of alternative and even more efficient methods for analysis of family-based case-control studies when parental genotype information is available on the case-control study participants. Based on these methods, we evaluate different family-based study designs by examining their relative efficiencies to each other and their efficiencies compared to a population-based case-control design of unrelated subjects. These comparisons reveal important design implications. Extensions of the methodologies for dealing with complex family studies are also discussed. *Genet. Epidemiol.* 28:138–156, 2005. © 2004 Wiley-Liss, Inc.

Key words: case-control studies; case-only designs; conditional likelihood; conditional logistic regression; family-based studies; gene-environment independence; gene-environment interactions; genetic epidemiology; matched case-control studies; population stratification.

Grant sponsor: National Cancer Institute; Grant number: CA-57030; Grant sponsor: National Institute of Environmental Health Sciences; Grant number: P30-ES09106.
*Correspondence to: N. Chatterjee, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, 6120 Executive Blvd, EPS Room 8038, Rockville, MD 20852. E-mail: chattern@mail.nih.gov
Received 11 June 2004; Accepted 7 July 2004
Published online 8 December 2004 in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20049

INTRODUCTION

The risks of many complex diseases are determined by a combination of the effects of genetic susceptibilities and environmental exposures. Advances in human genome research have thus led to epidemiologic investigations not only of the effects of genes alone, but also of their effects in combination with environmental exposures. The risk of a disease in relation to two exposures can be studied at various scales including statistical interaction (multiplicative or additive), joint effects, and the subgroup effect of one exposure within strata defined by the other exposures. Although the utility of a specific risk parameter depends on the scientific context and is sometimes a matter of debate, collectively various risk parameters of interest involving genetic and environmental exposures can be important for understanding biological and public health effects of the exposures, for targeting high-risk subjects for intervention, for individual risk prediction, and for enhancing the power to detect the association of the disease with one exposure (e.g., a gene) by selecting subjects to study based on the other exposure (e.g., the environment). Studies of genetic and environmental exposures together, whether designed to evaluate statistical

interactions or other parameters of scientific interest, can be cost-prohibitive due to the typical requirement of large sample sizes to achieve reasonable statistical power. Thus, efficient designs as well as efficient analytic methods that can reduce required sample size have become an important area of genetic-epidemiologic research.

Case-control study designs, in which diseased (cases) and non-diseased subjects (controls) are compared with respect to their exposure history, are now increasingly being used to study the role of genes and gene-environment interactions in the etiology of rare diseases. In population-based case-control designs, cases and controls are randomly selected from the diseased and non-diseased subjects that arise in an underlying population. Typically, the cases and controls in such designs are unrelated. In contrast, in family-based case-control designs, controls are selected from families of the cases. Both population- and family-based designs have advantages and disadvantages [Witte et al., 1999; Gauderman et al., 1999; Weinberg and Umbach, 2000]. While selection of population-based controls may be logistically more convenient, family-based designs can offer protection against spurious association induced by population stratification or admixture. Even when bias due to population stratification or admixture is not a concern, for efficiency reasons family-based designs may be preferred in studies of gene-environment interaction involving rare genetic variants [Witte et al., 1999; Gauderman, 2002].

The focus of this report is family-based designs where data on both genotype and environmental exposures are available on cases and related matching controls within families. Such designs may include, for example, designs where healthy siblings [Curtis, 1997; Spielman and Ewens, 1998] or cousins [Witte et al., 1999] of diseased subjects are selected as controls. Traditional analysis of such family-based studies involves conditional-logistic regression that restricts comparison of cases and controls to be within the family. The underlying theory of this approach relies on the conditional likelihood of the observed disease configuration data within matched-sets (family) given the risk factor information of the individual subjects. This method does not require any assumptions on the distribution of the risk factors in the underlying population.

In this article, we develop a new paradigm of conditional likelihoods for efficient analysis of family-based case-control studies when genetic susceptibility and environmental exposures can be assumed to be independently distributed of each other within families in the source population.

The efficiency advantage of methods that can exploit the gene-environment independence assumption was first observed in the context of population-based case-control studies. If genetic (G) and environmental (E) risk-factors can be assumed to be independently distributed in the underlying population, then the multiplicative interaction (also known as statistical interaction) between G and E for a rare disease can be estimated as the odds ratio between G and E among cases alone: the corresponding case-only estimate of interaction can be much more precise than the corresponding estimate of the interaction parameter from standard logistic regression analysis that involves both cases and controls but does not exploit the independence assumption [Piegorsch et al., 1994]. For discrete covariates, Umbach and Weinberg [1997] and then more generally Chatterjee and Carroll [2005] have developed efficient methods for estimating all of the parameters in a logistic regression model using data from both cases and controls and utilizing the *G-E* independence assumption.

The *G-E* independence assumption has been exploited also for the case-parent-trio design, an alternative family-based design that is known to be powerful for studying the effects of genes alone. The assumption was first used implicitly to show that the multiplicative interaction between G and E can be estimated from a case-parent-trio design that genotypes cases and their parents and determines environmental exposures of the cases [Schaid, 1999]. In particular, the likelihood based methods for analysis of such data rely on the assumption that conditional on parental genotypes, an individual's exposure status is independent of his/her genotype, a relatively weak independence assumption that is not affected by spurious association between genotype and exposure status in the general population that may be created due to hidden substructure [Umbach and Weinberg, 2000; Thomas, 2000].

Within the context of family-based studies, the choice of the case-control or the case-parents design for studies of gene-environment interaction depends on a number of different considerations [Weinberg and Umbach, 2000]. Besides various practical issues such as the availability of parents in the case-parent-trio design and availability of sibling/cousin for case-control designs, an important consideration is the relative

efficiency of these designs for estimation of various parameters of interest. Efficiency comparisons for estimation of the multiplicative interaction parameter have revealed that either the case-control design or the case-parent design can be superior depending on whether the effect of the gene under study is dominant or recessive, respectively [Witte et al. 1999; Gauderman, 2002]. It is worth noting that in all of these previous efficiency comparisons, the conditional likelihood method that is used for analysis of case-parent designs relies on the *G-E* independence assumption, while the traditional conditional logistic regression method used for analysis of case-control design does not exploit any such assumption.

In this article, we develop an alternative conditional likelihood framework for analyzing family-based case-control studies. Our approach relies on the assumption that genetic and environmental exposures are independently distributed within families. This family-level independence assumption, similar to the *G-E* independence assumption required for the case-parent design, is relatively weak in the sense that it is less likely to be affected by spurious association between *G* and *E* in the population. We show that when the underlying independence assumption is valid, the method can lead to major efficiency gains over traditional conditional logistic regression analysis of family-based case-control studies.

We also show that our conditional likelihood framework can be used to analyze data from a novel hybrid design that obtains genotype data of parents in addition to collecting genotype and environmental exposure data on cases and family-based controls. In particular, we show that a sibling-case-control design with parental genotype information, when analyzed using our new conditional likelihood approach, can be far superior to the sibling-case-control, case-parent, or population-based case-control design for estimation of different parameters of interest, and in a wide variety of situations. Various extensions of the methods for general family-based studies are also described.

BACKGROUND OF CONDITIONAL LOGISTIC REGRESSION (CLR)

MODEL AND NOTATION

For a major part of this article, we consider designs with 1:1 case-control matching. Later, we discuss how the methodology can be extended to more general types of family studies that may involve more than one case and/or control per family. We first describe a set of model assumptions that are required for traditional CLR analysis. Let (D_1, D_2), (G_1, G_2), and (E_1, E_2) denote the 0-1 disease indicators, genotypes, and environmental exposures for a pair of relatives. We assume that within a given family *F*, the risk of disease for two relatives is conditionally independent given their covariate information, and that the prospective risk model for the disease for the j^{th} ($j = 1$ or 2) relative is given by the logistic regression model

$$\text{pr}(D_j = 1 | G_j, E_j, F) = H\{\alpha_F + m(G_j, E_j; \beta)\}, \quad (1)$$

where $H(x) = \{1 + \exp(-x)\}^{-1}$ is the logistic distribution function and $m(\bullet)$ is a known but arbitrary function. Let $R_F(G_1, G_2, E_1, E_2)$ denote the joint distribution of (G_1, G_2) and (E_1, E_2) within family *F*. Standard CLR analysis allows $R_F(G_1, G_2, E_1, E_2)$ to be completely arbitrary.

There are two important features of the above model that need special attention. First, model (1) allows the family-specific intercept parameter α_F to account for potential heterogeneity in disease risk between different families. Such heterogeneity may arise, for example, if there are other sources of familial aggregation that cannot be explained by the genetic and environmental exposures under study.

Second, the model (1) allows the joint log-odds-ratio (log-relative-risk assuming rare disease) function $m(G, E, \beta)$ to be of very general form, so that it can include many different kinds of interaction models. The standard logistic model corresponds to $m(G, E, \beta) = \beta_G G + \beta_E E + \beta_{GE} G * E$, where $\exp(\beta_G)$ is the relative-risk (assuming rare disease) associated with the gene variant in the absence of the environmental exposure (the main effect of *G*), $\exp(\beta_E)$ is the relative-risk associated with the exposure in the absence of the gene-variant (the main effect of *E*), and $\exp(\beta_{GE})$ is the multiplicative interaction between *G* and *E*, which measures how the relative-risk associated with exposure changes with genotypes, or, equivalently, how the relative-risk associated with the gene variant changes with the exposure, with the changes being measured in the ratio scale. Many studies of gene and environment focus on estimation of the multiplicative interaction parameter β_{GE}, but estimation only of such statistical interaction may not necessarily contribute to an understanding of the biological or public health

effects of the two exposures [Thompson, 1991; Clayton and McKeigue, 2001]. Thus, when genetic and environmental factors are being studied together, it is important to consider modelling and estimation approaches that allow flexibility of estimation of various different parameters of interest. Examples of such parameters include the joint effect of the exposures G and E, the effect of E in sub-groups of subjects with different genetic exposures, the effect of G in sub-groups of subjects with different environmental exposures, and interaction between G and E at various different scales [Khoury, et al., 1993]. We will describe all of our methodologies in the general setting of model (1), which allows testing and estimation of all the risk-parameters of interest.

We assume M relative pairs are sampled into the study so that each pair has one diseased (case) and one non-diseased (control) subject. For the i^{th} such matched set, let D_{i0}, G_{i0}, and E_{i0} denote the disease status, genotype, and environmental exposure for the control and D_{i1}, G_{i1} and E_{i1} denote the corresponding values for the cases.

TRADITIONAL CONDITIONAL LIKELIHOOD

We now describe the conditional likelihood that forms the basis for traditional CLR analysis of family-based or other types of individually matched case-control studies. In the context of our model and notation, this conditional likelihood for the i^{th} matched set is given by

$$\begin{aligned} L_{i,\text{CLR}} =& \Pr(D_{i1}=1,\, D_{i0}=0 | D_{i1}+D_{i0}=1, \\ & G_{i1}, G_{i0}, E_{i1}, E_{i0}) \\ =& \frac{\exp\{m(G_{i1}, E_{i1}; \beta)\}}{\exp\{m(G_{i1}, E_{i1}; \beta)\} + \exp\{m(G_{i0}, E_{i0}; \beta)\}}. \end{aligned} \tag{2}$$

In (2), conditioning on the *set* event $D_{i1} + D_{i0} = 1$ reflects the constraint that by design the total number of cases in each matched set is exactly equal to one. For each matched set i, the conditional likelihood is formed based on the probability of the observed disease configuration for the members of the matched set, conditional on their risk factor information G and E and the ascertainment event $D_{i1} + D_{i0} = 1$. We observe that for studying the effect of genes alone using the sibling-case-control design, Spielman and Ewens [1998] previously proposed a Monte Carlo test-procedure, known as Sib-TDT (SDT), based on within-family permutation of genotypes. The CLR method can be viewed as an alternative likelihood-based analysis approach that is efficient as well as flexible in the sense that it allows both testing and estimation of risk parameters, can adjust for co-factors, and can be used to study gene-environment interaction.

There are several features of the above conditional likelihood that make the CLR analysis very flexible. First, computation of the CLR likelihood, as shown in the second line of formula (2), is free of the family-specific intercept parameters α_{F_i} and hence does not require any modelling assumptions about possible mechanisms of heterogeneity in disease risk between different families that cannot be explained by the genetic and environmental risk factors under study. Moreover, the likelihood in formula (2) is constructed based on probabilities that condition on all the risk factor information in a matched set and, hence, is free of any assumption about $R_F(G_1, G_2, E_1, E_2)$, the joint distribution of the risk factors in pairs of relatives. At the same time, however, the method, being distribution free, cannot exploit the gene-environment independence assumption when it is reasonable to do so.

PROPOSED METHODOLOGY

A NOVEL CONDITIONING PRINCIPLE

We propose to exploit an appropriate *G-E* independence assumption based on a conditional likelihood that does not condition on all of the genotype information of the individual subjects in a matched set. The straightforward choice for such a conditional likelihood is $\Pr(D_{i1}=1, D_{i0}=0, G_{i1}, G_{i0} | D_{i1}+D_{i0}=1, E_{i1}, E_{i0})$. However, a complication with such a conditional likelihood is that its computation depends on the joint genotype frequencies for pairs of relatives. In the presence of population substructure, genotype frequencies may vary across families and estimation of the regression parameters in the presence of family-specific genotype/allele frequency parameters is not possible without a strong modelling assumption about the distribution of the gene in different families. Since a major motivation of family-based designs is to avoid making this kind of assumption, we propose an alternative conditional likelihood that does not involve the genotype-frequency parameters and yet can exploit the *G-E* independence assumption. The most general

form of such a conditional likelihood is given by

$$L_{i,\text{general}} = \Pr(D_{i1} = 1, D_{i0} = 0, G_{i1}, G_{i0} | D_{i1} + D_{i0} = 1, \mathcal{G}_i, E_{i1}, E_{i0}), \quad (3)$$

where the conditioning event $\mathcal{G}_i$ denotes a variable that contains partial, but not all, information about the genotypes of the subjects in the matched set i, chosen in such a way that $L_{i,\text{general}}$ remains free of genotype-frequency parameters. Similar ideas for conditioning on partial genotype information to remove distributional assumption on genotypes have been previously used in the context of case-parent-trio studies of genetic association [Clayton, 1999; Cordell and Clayton, 2002; Rabinowitz and Laird, 2000]. In what follows, we will show how such an idea can be utilized for efficient analysis of gene-environment interaction in the context of case-control studies.

THE NEW CONDITIONING PARADIGM IN THE STANDARD CASE-CONTROL DESIGN

We now describe how the conditioning principle outlined above can be applied to analyze data from standard family-based case-control designs where genotype and exposure data are available only for selected cases and related matching controls. The required *G-E* independence assumption for this design is that the joint genotype and exposure status are independent for pairs of relatives within each family in the source population. Formally speaking, this assumption for a given family F implies that the joint genotype and exposure distribution $R_F(G_1, G_2, E_1, E_2)$ can be expressed as

$$R_F(G_1, G_2, E_1, E_2) = Q_F(G_1, G_2) \times V_F(E_1, E_2), \quad (4)$$

where Q_F and V_F are the family specific distributions of (G_1, G_2) and (E_1, E_2), respectively. We also assume that the joint genotype distribution for any pair of relatives is symmetric, that is $Q_F(g_1, g_2) = Q_F(g_2, g_1)$, an assumption that automatically holds under the Mendelian law of inheritance within families. Hereafter, we will describe the assumption stated in (4) as "Type-I Independence" to distinguish it from an alternative independence assumption that we will use later for case-control designs with parental genotype information.

Observe that independence assumption (4) is much weaker than the population-based *G-E* independence assumption that is required for the case-only design. The population-based independence assumption, for example, may be violated due to the effects of a hidden population substructure [Umbach and Weinberg, 2000] or the influence of family history, a factor that is clearly related to susceptibility genes, on lifestyle factors such as smoking [Thomas, 2000]. Since related subjects share both ethnic and family history background, the within-family independence assumption (4) is much less likely to be affected by spurious association due to these factors. In particular, the assumption is the weakest for the sibling-case-control design, because siblings share ethnic and family history background. Cousins, on the other hand, only partially share these factors and thus the required assumption for the cousin case-control design is stronger. Possible ways for further relaxing this assumption based on a conditional independence model will be discussed later.

In the setting of the standard case-control design, we propose the conditioning event $\mathcal{G}_i$ in the likelihood (3) to be $\mathcal{G}_i^S$, the *set* of genotypes that is observed in the i^{th} matched pair. For matched pairs where observed genotypes are discordant, $\mathcal{G}_i$ contains the information on the two different types of genotypes that are observed for that pair, but does not specify the individual genotype of the case (G_{i1}) and the control (G_{i0}). In the Appendix, we show that under a rare disease assumption, with this definition of $\mathcal{G}_i$ the proposed conditional likelihood (3) can be computed as

$$L_{i,\text{CC}} = \frac{\exp\{m(G_{i1}, E_{i1}; \beta)\}}{\sum_{j=0}^{1} \left[\exp\{m(G_{ij}, E_{i1}; \beta)\} + \exp\{m(G_{ij}, E_{i0}; \beta)\}\right]}. \quad (5)$$

The sum in the denominator of $L_{i,\text{CC}}$ essentially constitutes four subjects corresponding to the four genotype-exposure configurations: (G_{i0}, E_{i0}), (G_{i1}, E_{i1}), (G_{i0}, E_{i1}), and (G_{i1}, E_{i0}). Thus, $L_{i,\text{CC}}$ is the same as the standard conditional likelihood for a 1:3-matched design, where the two additional subjects with genotype-exposure configuration (G_{i0}, E_{i1}) and (G_{i1}, E_{i0}) can be viewed as "pseudo" family members obtained by exchanging the genotypes of the observed family members: under the *G-E* independence assumption, such "pseudo" subjects are as equally likely to appear in a family as the observed subjects in that family. In this spirit, the proposed methodology has an intriguing similarity with the conditional logistic

regression analysis of case-parent trio designs [Self et al., 1991], which is also based on pseudo-sibs for the observed case that could have been observed given the genotype of the parents.

Several other observations can be made from the final form of the conditional likelihood (5). First, similar to the standard conditional likelihood, $L_{i,\mathrm{CLR}}$ in (2), $L_{i,\mathrm{CC}}$ is free of the family-specific intercept parameters α_{F_i}. Second, although $L_{i,\mathrm{CC}}$ relies on the assumption of independence between G and E within each family, it is otherwise quite flexible in the sense that it does not depend on any assumption about $V_F(E_1, E_2)$, the family-specific distributions of joint exposure status for two relatives. Finally, because of the standard CLR form, estimates of the regression parameters β that maximize $L_{i,\mathrm{CC}}$, as well as corresponding asymptotic variance estimates, can be obtained by using standard and widely available CLR software.

In the Appendix, we derive standard errors for the estimates of β for a general class of designs that include all the designs described in this report. The relevant formula is (A.5). Also, in the Appendix we show that the proposed method is asymptotically at least as efficient as standard CLR. Indeed, what we show is that the asymptotic covariance matrix of estimates of β that maximize $L_{i,\mathrm{CC}}$ can be written as

$$V = I^{-1}(\beta) = \{I_1(\beta) + I_2(\beta)\}^{-1},$$

where $I_1(\beta)$ is the information matrix corresponding to the standard conditional likelihood (2) and $I_2(\beta)$ is a non-negative definite matrix. This automatically proves that $\{I_1(\beta) + I_2(\beta)\}^{-1} \leq \{I_1(\beta)\}^{-1}$ in the sense $\{I_1(\beta) + I_2(\beta)\}^{-1} - \{I_1(\beta)\}^{-1}$ is always a non-positive-definite matrix, thus implying that the proposed method is asymptotically at least as efficient as standard CLR.

Finally, we observe that the within family *G-E* independence assumption can be also exploited to construct powerful permutation-based methods for testing a global null hypothesis of no association. The pivot statistics could be given by any measure of distance, such as the sum of square differences, between the joint genotype and exposure frequencies of the cases and those of the controls. The permutation distribution of the statistics can be then generated by randomly switching G and E status within matched pairs of cases and controls. Unlike standard permutation tests, where covariates are permuted together, under the *G-E* independence assumption the two type of exposures should be permuted independently of one another.

THE NEW CONDITIONING PARADIGM IN CASE-CONTROL DESIGN WITH PARENTAL GENOTYPE DATA

For simplicity of notation, we will describe the proposed method in the context of a case-sibling-control design with parental genotype information. The method easily extends to alternative types of matched case-control designs as long as genotype data are available for parents of both cases and controls.

We assume M matched case-control sibling-pairs are sampled into a study. We use the same data structure and notation that we have introduced earlier for the standard family-based case-control designs. In addition, we define $\mathcal{G}_i^P = (G_{iM}, G_{iF})$ to be the parental (mother and father) genotype data for the i^{th} matched pair. The key assumption we exploit in this design setting is that genotype and exposure status for pairs of relatives in the source population are independently distributed conditional on their parental genotype information. Thus, if (G_1, G_2) and (E_1, E_2) denote the joint genotype and exposure status for a pair of siblings and $\mathcal{G}^P$ denotes their parental genotype information, the required independence assumption can be stated formally as

$$\mathrm{pr}(G_1, G_2, E_1, E_2|\mathcal{G}^P) = \mathrm{pr}(G_1, G_2|\mathcal{G}^P) \times \mathrm{pr}(E_1, E_2|\mathcal{G}^P). \tag{6}$$

Moreover, assuming a Mendelian mode of inheritance given parental genotypes, we can write $\mathrm{pr}(G_1, G_2|\mathcal{G}^P) = \mathrm{pr}(G_1|\mathcal{G}^P) \times \mathrm{pr}(G_2|\mathcal{G}^P)$. The family-based independence assumption stated in formula (6) is very weak in the sense that it is robust to the effects of various factors such as the presence of hidden population sub-structure or the influence of family history on lifestyle-related exposures. We will describe the assumption stated in (6) as "Type-II Independence".

In the setting described above, we propose to use the conditioning event $(\mathcal{G}_i)$ in the general conditional likelihood (3) to be the *parental* genotype information $(\mathcal{G}_i^P)$. We observe that $\mathcal{G}_i^P$ is a larger event than $\mathcal{G}_i^S$, the *set* genotype information in a case-control pair, in the sense that $\mathcal{G}_i^P$ contain the information of all possible values for $\mathcal{G}_i^S$. Thus, it is expected that when parental

genotype data are available, conditioning on $\mathcal{G}_i^P$ will be more efficient than the conditioning on $\mathcal{G}_i^S$. The general conditional likelihood (3) with $\mathcal{G}_i = \mathcal{G}_i^P$ can be computed as

$$L_{i,\mathrm{CCGP}} = \frac{\exp\{m(G_{i1}, E_{i1}; \beta)\}\mathrm{pr}\left(G_{i1}|\mathcal{G}_i^P\right)\mathrm{pr}\left(G_{i0}|\mathcal{G}_i^P\right)}{\sum_{G'_{i1} \in H_{\mathcal{G}_i^P}} \exp\{m(G'_{i1}, E_{i1}; \beta)\}\mathrm{pr}\left(G'_{i1}|\mathcal{G}_i^P\right) + \sum_{G'_{i0} \in H_{\mathcal{G}_i^P}} \exp\{m(G'_{i0}, E_{i0}; \beta)\}\mathrm{pr}\left(G'_{i0}|\mathcal{G}_i^P\right)}, \quad (7)$$

where $H_{\mathcal{G}_i^P}$ denotes all possible offspring genotypes associated with the parental genotypes G_i^P. The quantities $\mathrm{pr}\left(G_{ij}|\mathcal{G}_i^P\right)$, the genotype probability of an offspring given the genotype of the parents, can be computed as fixed constants assuming a standard Mendelian mode of inheritance within family. Derivation of formula (7) follows assuming rare disease and calculations similar to those for the derivation of formula (5), and is given in the Appendix. In terms of computation, estimates of β maximizing $L_{i,\mathrm{CCGP}}$ and associated standard errors can be obtained using standard CLR software that allows incorporation of offset terms. The general form of the covariance matrix of the parameter estimates is given in equation (A.5) in the Appendix.

There are connections between L_{CCGP} and the traditional conditional-likelihood for case-parents-trio (CPT) data [Self et al., 1991; Schaid, 1999], which is given by

$$L_{i,\mathrm{CPT}} = \frac{\exp\{m(G_{i1}, E_{i1}, \beta)\}\mathrm{pr}\left(G_{i1}|\mathcal{G}_i^P\right)}{\sum_{G'_{i1} \in H_{\mathcal{G}_i^P}} \exp\{m(G'_{i1}, E_{i1}, \beta)\}\mathrm{pr}\left(G'_{i1}|\mathcal{G}_i^P\right)}. \quad (8)$$

The numerators of both $L_{i,\mathrm{CCGP}}$ and $L_{i,\mathrm{CPT}}$ correspond to the i^{th} case, with the expressions being the same up to a constant term. However, there are differences in the denominator. While the denominator of $L_{i,\mathrm{CPT}}$ consists of all possible offspring the i^{th} pair of parents could have had with the offsprings' environmental exposure the same as that of the observed case (E_{i1}), the denominator of $L_{i,\mathrm{CCGP}}$ consists of those of $L_{i,\mathrm{CPT}}$ plus all possible offspring the i^{th} pair of parents could have had with the offsprings' environmental exposure the same as that of the observed control (E_{i0}).

In recent years, various extensions of $L_{i,\mathrm{CPT}}$ have been developed for utilizing data on multiple offspring in nuclear families [Clayton, 1999; Cordell and Clayton, 2002; Cordell et al., 2004; Kraft et al., 2004]. All of these methods, however, condition on the exact phenotype (disease status) of the individual offsprings. As a result, in these methods, unaffected subjects (controls) are only indirectly informative, in the sense that they can be utilized to infer missing parental genotype data, but once the parental genotype information is available/inferred, the unaffected subjects are ignored in the respective conditional likelihoods. In contrast, $L_{i,\mathrm{CCGP}}$ conditions only on the *set* of phenotypes defined by the event $D_1 + D_0 = 1$, instead of the individual phenotypes D_1 and D_0 themselves. In this approach, the unaffected offspring remain informative even when complete parental genotype information is available. The environmental exposure status (E_0) of the unaffected subject allows estimation of β_E, the main effect parameter associated with E, which cannot be estimated from $L_{i,\mathrm{CPT}}$. Moreover, incorporation of the unaffected subjects leads to major increase in efficiency for estimation of the multiplicative interaction parameter (β_{GE}) and other related quantities (see Tables I and II). It is, however, important to note that $L_{i,\mathrm{CCGP}}$, similar to $L_{i,\mathrm{CC}}$, requires the assumption that the selection of a case-control pair of relatives does not depend on the individual G and E status of the relatives [Hsu et al., 2000].

SIMULATION STUDIES INVOLVING DIFFERENT DESIGNS AND ANALYTIC METHODS FOR FAMILY-BASED CASE-CONTROL STUDIES

In this section, we report simulation studies of the relative efficiency of different study designs and analytic methods for estimation of various risk-parameters of interest using data from nuclear families. In particular, we considered three designs: (A) the Sibling-Case-Control (SCC) design with G and E available on the matched cases and controls; (B) the Case-parent-trio (CPT) design with G available on cases and their parents and E available on the cases; and (C) the Sibling-Case-Control design with genotyped parents (SCCGP). The analytic methods we compared are: (1) Traditional conditional likelihood ($L_{i,\mathrm{CLR}}$) for design (A); (2) the proposed conditional likelihood

TABLE I. Dominant Gene: Bias and efficiencies of alternative family-based designs[a] and analytic methods[m] for evaluation of different risk parameters

				SCC[d2]					
		CPT[d1]: L^{m1}_{CPT}		L^{m2}_{CLR}		L^{m3}_{CC}		SCCGP[d3]: L^{m4}_{CCGP}	
$\{p_G, p_E\}$[b] $\{\beta_G, \beta_E\}$[c]	Risk-parameters[d]	Bias[e]	RE[f]	Bias[e]	RE[f]	Bias[e]	RE[f]	Bias[e]	RE[f]
0.01, 0.2	OR(G | E=1)	0.09	1.13	−0.01	1.62	0.03	2.30	0.08	3.49
log(7), log(1.3)	OR(E | G=1)	NA	NA	−0.02	2.89	0.09	5.52	0.10	5.67
	MI_{GE}	0.18	0.81	−0.02	2.90	0.11	5.52	0.13	5.67
	AI_{GE}	NA	NA	−0.02	2.90	0.09	5.56	0.11	5.64
0.2, 0.2	OR(G | E=1)	0.05	1.04	0.00	0.82	0.02	1.04	0.04	1.52
log(1.3), log(1.3)	OR(E | G=1)	NA	NA	0.00	0.93	0.02	1.19	0.03	1.40
	MI_{GE}	0.06	0.94	−0.00	1.05	0.02	1.34	0.04	1.64
	AI_{GE}	NA	NA	0.00	1.05	0.02	1.40	0.04	1.72
0.01, 0.5	OR(G | E=1)	0.09	0.82	−0.01	0.66	0.02	0.78	0.08	1.38
log(7), log(1.12)	OR(E | G=1)	NA	NA	−0.01	2.49	0.09	3.89	0.10	4.18
	MI_{GE}	0.19	0.70	−0.01	2.39	0.10	3.58	0.12	3.83
	AI_{GE}	NA	NA	−0.01	2.48	0.09	3.91	0.11	4.18
0.2, 0.5	OR(G | E=1)	0.04	0.79	−0.00	0.56	0.01	0.63	0.03	1.05
log(1.3), log(1.12)	OR(E | G=1)	NA	NA	0.00	0.94	0.02	1.14	0.02	1.34
	MI_{GE}	0.05	0.79	0.00	0.96	0.02	1.16	0.03	1.37
	AI_{GE}	NA	NA	0.00	0.94	0.02	1.17	0.03	1.38

[a]Designs: [d1]Case-parents trio, [d2]Sibling case-control, and [d3]Sibling case-control w. genotyped parents. Methods[m]: Conditional-likelihoods described in formulae [m1](8), [m2](2), [m3](5), and [m4](7).
[b]Genotype (G=*Aa*/*aa*) and exposure (E=1) frequencies.
[c]True values for main effects of G and E.
[d]OR(G | E=1): OR for G among subjects with E=1; OR(E | G=1): OR for E among subjects with G=1; MI_{GE}; multiplicative-interaction; AI_{GE}: additive interaction.
[e]Relative bias evaluated as (true value – mean estimated value)/true value.
[f]Relative efficiencies compared to a population-based case-control design with the same number of cases and 1:1 case-control ratio.

($L_{i,\text{CC}}$) for design (A); (3) the efficient conditional likelihood ($L_{i,\text{CPT}}$) method for analysis of designs (B) [Self et al., 1991; Schaid, 1999]; and (4) the proposed conditional likelihood ($L_{i,\text{CCGP}}$) for design (C).

THE SIMULATION DESIGN

We assumed that the gene variant of interest is a bi-allelic locus with the wild and variant-type alleles being denoted by *A* and *a*, respectively. We considered two distinct settings of interest, one for rare variants and the other for common variants. Within each setting, we considered dominant and recessive models for the effect of the gene-variant. We also assumed a binary environmental exposure and considered two scenarios, one involving a common exposure and the other involving a rare exposure.

We simulated data for nuclear families consisting of two siblings and their parents using the following setup. We simulated a family-specific allele frequency parameter to allow for population-substructure. For each family (F), we simulated an allele frequency parameter (θ_F) by first generating a random variable u_F from the normal distribution with mean parameter μ and variance σ^2 and then transforming u_F to the 0-1 scale as $\theta_F = \exp(u_F)/\{1+\exp(u_F)\}$. We chose the variance parameter σ^2 to be 0.5 so that the $\pm 2\sigma$ limit of this distribution corresponds to approximately 15-fold variation in allele frequency across different families. We chose the mean parameter μ in such a way that the marginal probability of the genotype variant of interest (*Aa* or *aa* for the dominant model and *aa* for the recessive model) in the underlying population is fixed at 0.01 for the setting of a rare variant and 0.2 for the setting of a common variant. Given the allele frequency parameter θ_F for a family, we generated the genotype data for the parents assuming Hardy-Weinberg-Equilibrium and that the parents are

TABLE II. Recessive gene: bias and efficiencies of alternative family-based designs[a] and analytic methods[m] for evaluation of different risk parameters

				SCC[d2]					
		CPT[d1]: L^{m1}_{CPT}		L^{m2}_{CLR}		L^{m3}_{CC}		SCCGP[d3]: L^{m4}_{CCGP}	
$\{p_G, p_E\}$[b] $\{\beta_G, \beta_E\}$[c]	Risk-parameters[d]	Bias[e]	RE[f]	Bias[e]	RE[f]	Bias[e]	RE[f]	Bias[e]	RE[f]
0.01, 0.2	OR(G | E=1)	0.11	2.81	−0.02	1.69	0.04	2.74	0.10	5.43
log(7), log(1.3)	OR(E | G=1)	NA	NA	−0.03	2.49	0.12	4.62	0.14	5.74
	MI_{GE}	0.21	2.04	−0.04	2.49	0.14	4.09	0.18	5.12
	AI_{GE}	NA	NA	−0.03	2.51	0.12	4.67	0.15	5.78
0.2, 0.2	OR(G | E=1)	0.05	1.28	0.00	0.69	0.02	0.95	0.04	1.73
log(1.3), log(1.3)	OR(E | G=1)	NA	NA	0.00	0.92	0.02	1.26	0.03	1.45
	MI_{GE}	0.06	1.13	−0.00	0.96	0.03	1.40	0.04	1.80
	AI_{GE}	NA	NA	0.00	0.93	0.02	1.39	0.04	1.86
0.01, 0.5	OR(G | E=1)	0.10	2.06	−0.01	0.90	0.02	1.14	0.09	2.99
log(7), log(1.12)	OR(E | G=1)	NA	NA	−0.01	2.26	0.11	4.33	0.13	5.26
	MI_{GE}	0.20	1.77	−0.01	2.26	0.13	3.93	0.15	4.83
	AI_{GE}	NA	NA	−0.01	2.26	0.11	4.33	0.14	5.28
0.2, 0.5	OR(G | E=1)	0.04	1.14	−0.00	0.59	0.01	0.70	0.04	1.29
log(1.3), log(1.12)	OR(E | G=1)	NA	NA	0.00	0.79	0.02	1.05	0.02	1.13
	MI_{GE}	0.05	0.80	0.00	0.73	0.02	0.95	0.04	1.13
	AI_{GE}	NA	NA	0.00	0.72	0.02	0.98	0.03	1.17

[a]Designs: [d1]Case-parents trio, [d2]Sibling case-control, and [d3]Sibling case-control w. genotyped parents. Methods[m]: Conditional-likelihoods described in formulae [m1](8), [m2](2), [m3](5), and [m4](7).
[b]Genotype (G=*aa*) and exposure (E=1) frequencies.
[c]True values for main effects of G and E.
[d]OR(G | E=1): OR for G among subjects with E=1; OR(E | G=1): OR for E among subjects with G=1; MI_{GE}; multiplicative-interaction; AI_{GE}: Additive interaction.
[e]Relative bias evaluated as (true value−mean estimated value)/true value.
[f]Relative efficiencies compared to a population-based case-control design with the same number of cases and 1:1 case-control ratio.

independent. Given the genotype of the parents, we generated the genotypes for a pair of siblings based on a standard Mendelian mode of inheritance. We generated the environmental exposures for a pair of siblings by first generating a pair of correlated random variables (E_1^*, E_2^*) from a bivariate normal distribution with marginal means zero, marginal variances one and a correlation parameter ρ. We then dichotomized E_1^* and E_2^* into two binary 0-1 exposure variables E_1 and E_2 so that the marginal probability of exposure $(E = 1)$ for the underlying population is 0.2 for the setting of a rare exposure and 0.5 for the setting of a common exposure. In our basic simulation setting (Tables I and II), we fixed the correlation parameter $\rho = 0.3$ so that it represents only a modest correlation between the environmental exposures for a pair of siblings. Later (Figs. 1 and 2), we explore the effect of varying ρ on the efficiencies of different designs and analytic methods.

We simulated the family-specific intercept term (α_F) to allow for heterogeneity in disease-risk between families that cannot be accounted for by G and E. For a given family F, we generated α_F from the normal distribution with mean α and variance τ^2. We chose $\tau^2 = 1$ so that the $\pm 2\sigma$ limit of this distribution corresponds to an approximately 50-fold variation in disease risk between families due to unknown factors. We fixed the mean parameter α at different values for different settings so that the marginal probability of the disease in the population, $\mathrm{pr}(D = 1)$, is always fixed at 0.01. Given α_F, we generated the disease outcome for each sibling, independent of the other, using the logistic regression model

$$\mathrm{pr}(D_j = 1 | G_j, E_j, \alpha_F) = \frac{\exp\{\alpha_F + \beta_G f(G) + \beta_E E + \beta_{GE} f(G) * E\}}{1 + \exp\{\alpha_F + \beta_G f(G) + \beta_E E + \beta_{GE} f(G) * E\}}, \quad (9)$$

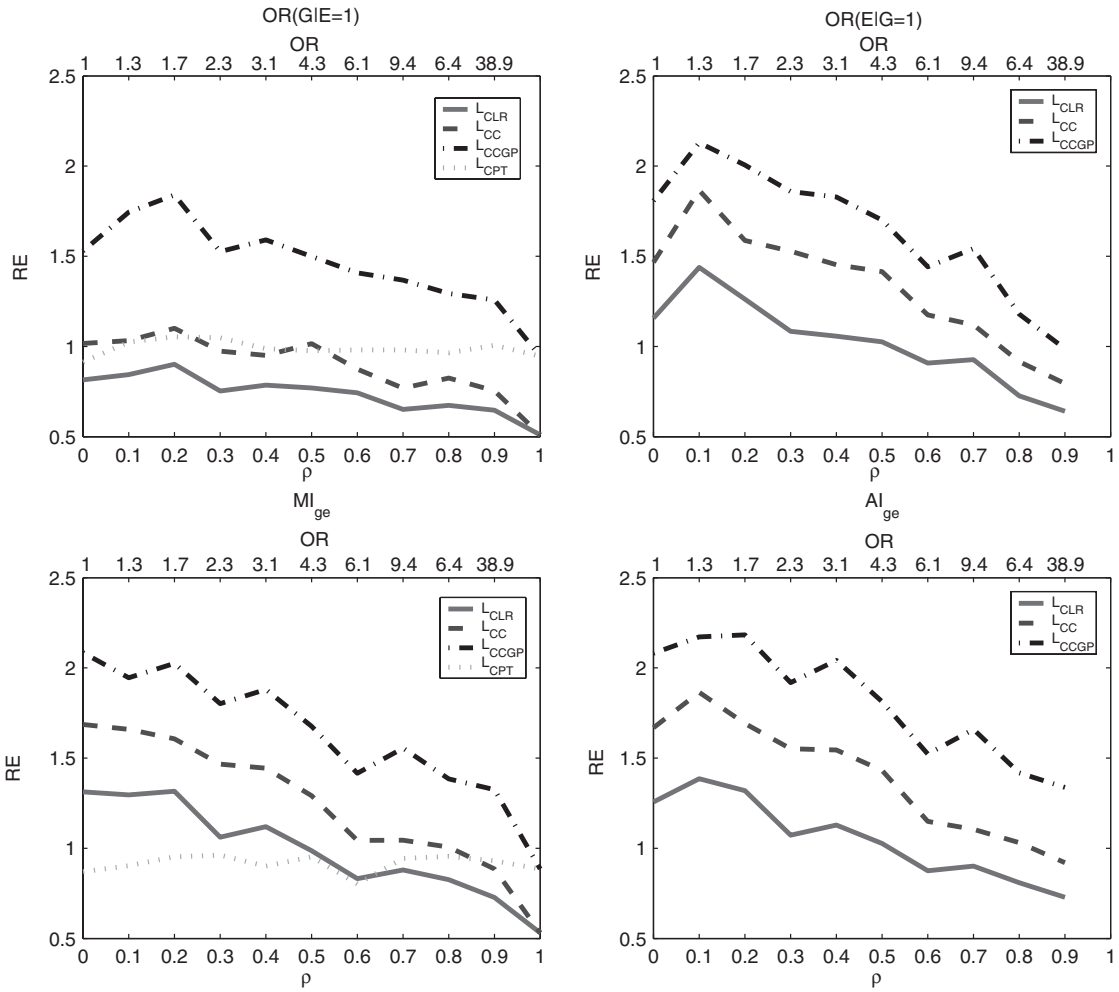

Fig. 1. Dominant gene: Relative efficiency (RE) of alternative family-based designs and analytic methods as a function of sibling-correlation (ρ) in exposure (E). The top axis shows the correlation in the binary OR scale. The methods compared are $\mathcal{L}_{\mathrm{CLR}}$ (solid line), $\mathcal{L}_{\mathrm{CC}}$ (dashed line), $\mathcal{L}_{\mathrm{CPT}}$ (dotted line), and $\mathcal{L}_{\mathrm{CCGP}}$ (dashed/dotted line). MI_{GE} and AI_{GE} indicate multiplicative and additive interaction parameters, respectively. Values of the relative efficiency are evaluated based on 500 simulations. In each simulation, data for different designs are generated based on 5,000 cases.

where $f(G)$ is a binary 0-1 function reflecting the mode of effect of the gene: $f(Aa/aa) = 1$ for dominant and $f(aa) = 1$ for recessive. In the basic simulation setting (Tables I and II), we chose the main effect parameters β_G to be $\log(7)$ when $\mathrm{pr}\{f(G) = 1\} = 0.01$ and $\log(1.3)$ when $\mathrm{pr}\{f(G) = 1\} = 0.2$, so that the two settings corresponds to a high-penetrance rare variant and a low-penetrance common variant, respectively. Similarly, we chose β_E to be $\log(1.3)$ when $\mathrm{pr}(E = 1) = 0.2$ and $\log(1.12)$ when $\mathrm{pr}(E = 1) = 0.5$, so that the main effect of E is stronger when the exposure is rare. We fixed β_{GE} to be $\log(3)$, which corresponds to a strong multiplicative interaction between G and E.

Following this scheme, we first generated data for a large number of randomly sampled nuclear families. Treating these randomly selected families as the underlying population, we then selected 5,000 families with one diseased and one non-diseased sibling. During analysis of data from each design, we only retained the appropriate genotype and environmental exposure information for that design and discarded the rest of the information.

In the above simulation setting, we used two types of random effects, namely u_F and α_F, to generate variations in genotype frequencies and disease-risk, respectively, across families. We chose these random effects to be uncorrelated with each other so that there is no population-level association between genetic exposure and disease risk that cannot be explained by the direct effect of the gene on risk of the disease other than

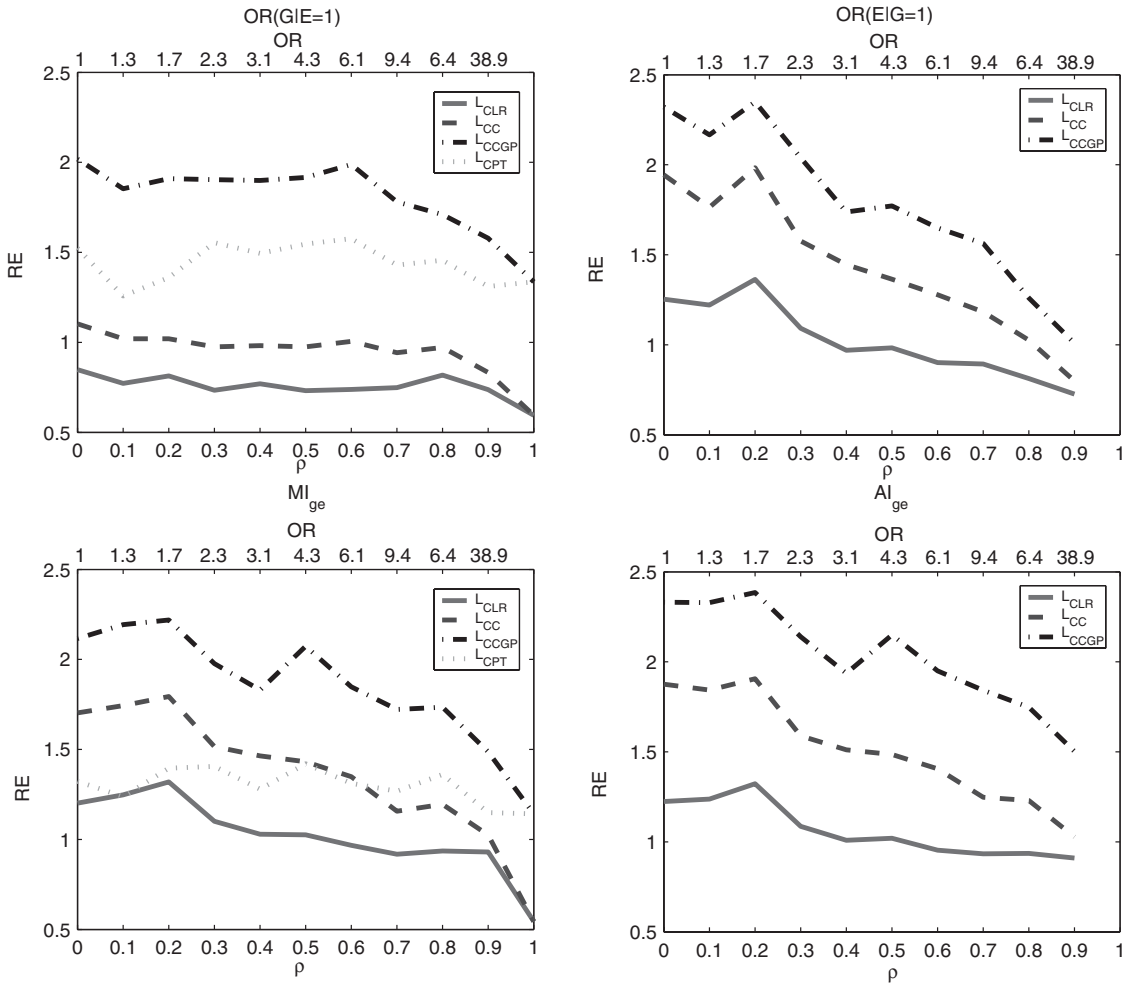

Fig. 2. Recessive gene: Relative efficiency (RE) of alternative family-based designs and analytic methods as a function of sibling-correlation (ρ) in exposure (E). The top axis shows the correlation in the binary OR scale. The methods compared are $\mathcal{L}_{CLR}$ (solid line), $\mathcal{L}_{CC}$ (dashed line), $\mathcal{L}_{CPT}$ (dotted line), and $\mathcal{L}_{CCGP}$ (dashed/dotted line). MI_{GE} and AI_{GE} indicate multiplicative and additive interaction parameters, respectively. Values of the relative efficiency are evaluated based on 500 simulations. In each simulation, data for different designs are generated based on 5,000 cases.

through the parameters β_G and β_{GE} in the risk model (9). Thus, in this setting, bias due to population-stratification not being a concern, a case-control study based on unrelated subjects is an alternative valid design. We chose the population-based case-control design (PCC), analyzed with standard logistic regression, to be the common reference point for evaluating the relative efficiencies of various family-based designs and analytic methods. To simulate data for this design, we first generated data for a large number of randomly sampled nuclear families and then selected 5,000 cases and 5,000 controls from 10,000 independent families.

For evaluating the efficiencies of different designs and methods, we considered four different parameters of epidemiologic interest: (1) $\mathrm{OR}(G|E=1)=\exp(\beta_G+\beta_{GE})$: the odds ratio associated with the gene variant among subjects with environmental exposure ($E=1$); (2) $\mathrm{OR}(E|G=1)=\exp(\beta_E+\beta_{GE})$: the odds ratio associated with the environmental exposure among subjects with a variant genotype ($G=1$); (3) $\mathrm{MI}_{GE}=\exp(\beta_{GE})$: the multiplicative interaction between G and E; and (4) $\mathrm{AI}_{GE}=\exp(\beta_G+\beta_E+\beta_{GE})-\exp(\beta_G)-\exp(\beta_E)+1$: the additive interaction between G and E [Khoury et al., 1993]. All of the designs except the case-parent-trio design yield an estimate of all of the four types of association/interaction parameters; the case-parent trio design cannot estimate the main effect of E (β_E) and hence also cannot estimate $\mathrm{OR}(E|G=1)$ and AI_{GE}.

Strictly speaking, estimates for a particular type of association/interaction parameter from

different designs are not directly comparable due to differences in scale of measurement. That is, while the model for analyzing the CPT design is defined in terms of within family relative risk parameters, those for family-based and population-based case-control designs are defined in terms of within- and between-family odds-ratio parameters, respectively. In spite of such differences, one common goal of all of the designs is to test for hypotheses about different types of association/interaction parameters. Thus, we evaluated different designs based on their relative powers for rejecting the null hypothesis about different association/interaction parameters in the respective underlying scales. For each type of association/interaction parameter of interest (θ), we evaluated the quantity $\tau = \log(\hat{\theta})/\text{sd}\{\log(\hat{\theta})\}$, where $\log(\hat{\theta})$ and $\text{sd}\{\log(\hat{\theta})\}$ are the empirical mean and the standard error of the estimate of θ from a given design over different simulated data sets. The ratio of τ^2 for two designs estimates the asymptotic relative efficiency of the two designs, i.e., the inverse-ratio of the sample sizes required by the two designs to reject the null hypothesis of no association/interaction, i.e., $\log(\theta) = 0$. For rare diseases, such as the one considered in our simulation setting, the differences between the mean estimates of parameters $\log(\hat{\theta})$ from different designs are small and thus the relative efficiencies of different designs are mostly determined by the precision of parameter estimates $1/\text{var}\{\log(\hat{\theta})\}$.

RESULTS: BIAS AND EFFICIENCY

Table I (dominant gene) and Table II (recessive gene) show the results of simulation experiments with fixed sets of parameter values. We make several key observations from Tables I and II, as follows. For the scenario of ''low penetrance common variant'' ($\beta_G = \log(1.3), \text{pr}\{f(G) = 1\} = 0.2$), all of the proposed analytic methods had negligible percentage bias in estimating the ''true'' parameters of the underlying disease-risk model. For the scenario of ''high-penetrance rare variant,'' the novel methods produced noticeable, but modest ($\leq 15\%$), bias in parameter estimates. The bias likely arises due to the rare disease approximation, because in this setting the risk of disease for subjects with both the genetic and the environmental exposure was as high as 30%. Comparison of the traditional (L_{CLR}) and the proposed method (L_{CC}) of analyzing the SCC design shows the major efficiency advantages of the latter approach for all of the four different parameters. Comparison of the CPT design and the SCC design shows that the latter design, when analyzed using our new method (L_{CC}), was superior for estimation of the multiplicative interaction term. For estimation of $\text{OR}(G|E = 1)$, however, the CPT design was generally more efficient. The SCCGP design, when analyzed using our approach (L_{CCGP}), had the highest efficiency among all of the designs for all of the four different association/interaction parameters. The efficiencies of all of the family-based designs relative to the PCC design decreased as either the genetic or the environmental exposure becomes more common.

Figures 1 (for dominant gene) and 2 (for recessive gene) show the effect of varying the parameter ρ that determines the correlation between the exposure variables in a pair of siblings. For these graphs, we chose other parameter values of the simulation in such a way so that they reflect an intermediate situation between the ''high-penetrant rare gene'' and ''the low-penetrant common gene'' scenarios we considered for Table I and Table II. In particular, we fixed $\Pr(G = 1) = 0.1$, $\Pr(E = 1) = 0.3$, $\beta_G = \log(1.6)$, $\beta_E = \log(1.12)$, and $\beta_{GE} = \log(3)$.

Overall, for all of the four types of parameters, the efficiencies of the SCC and SCCGP designs, relative to the PCC design, decreased as ρ increased. The efficiency of the CPT design did not depend on ρ as this design does not involve the sibling controls. Comparison of the traditional (L_{CLR}) and novel conditional likelihood method (L_{CC}) for analyzing the SCC design shows that the efficiency advantage of the latter method remained fairly constant over a wide range of value of ρ: only at very high values of ρ did the difference between the two method starts diminishing. When $\rho = 1$, which corresponds to two siblings having identical exposures, the two methods are identical and hence had the same efficiency. The efficiency of the SCCGP design also remained substantially higher than all the other designs except for extremely high values of ρ. At $\rho = 1$, the SCCGP and CPT design are identical. At $\rho = 1$, none of the family-based designs can estimate the association parameter $OR(E|G = 1)$ and the additive interaction parameter AI_{GE}.

Inspection of the efficiencies of the family-based designs relative to the PCC design suggests that in our simulation setting, where the gene-variant is moderately common, i.e., $\Pr(G = 1) = 0.1$, the

SCC design, when analyzed using the traditional method (L_{CLR}), had a lower efficiency than the PCC design for high values of ρ ($\rho > 0.5$). In contrast, when analyzed with the novel method (L_{CC}), the efficiency of the SCC design remained higher in a wider range of ρ. The efficiency of the SCCGP design remained higher than that for the PCC design for almost all values of ρ.

RESULTS: TYPE-I ERROR RATE

Under a global null hypothesis of no association of the disease with either G or E, the conditional probabilities $\Pr(D_{i1}=1, D_{i0}=0, G_{i1}, G_{i0} | D_{i1}+D_{i0}=1, G_i, E_{i1}, E_{i0})$ become free of the intercept parameters irrespective of whether the disease is rare or not. Thus, the tests based on our proposed likelihood will be valid under this global null hypothesis whether the disease is rare or not. We evaluated the empirical type-I error rates of the proposed methods for testing weaker null hypotheses that specify only certain parameters of interest to be null, but leave the other parameters unspecified.

We consider the simulation setting of a high-penetrant dominant rare gene, a scenario where we had observed modest bias in parameter estimation due to violation of the rare disease assumption. We simulated data under four types of null hypotheses corresponding to (1) $\mathrm{MI}_{GE}=1$, (2) $\mathrm{OR}(G|E=1)=1$, (3) $\mathrm{OR}(E|G=1)$, and (4) $\mathrm{AI}_{GE}=0$. For generating data under (1), we chose β_G to be $\log(7)$ and β_E to be $\log(1.3)$ as before, but set $\beta_{GE}=0$. For generating data under (2), we chose $\beta_{GE}=-\beta_G=-\log(7)$ and $\beta_E=\log(1.3)$. Similarly, for generating data under (3), we chose $\beta_{GE}=-\beta_E=-\log(1.3)$ and $\beta_G=\log(7)$. Finally, for generating data under (4), we chose $\beta_G=\log(7)$ and $\beta_E=\log(1.3)$ and then solved for that value of β_{GE} so that $\exp(\beta_G+\beta_E+\beta_{GE})-\exp(\beta_G)-\exp(\beta_E)+1=0$. Table III shows the empirical type-I error rates of the Wald tests (5% significance level) associated with the likelihoods L_{CC} and L_{CCGP}. We observe that for each type of null hypotheses, the test procedures maintained the nominal α -level very well. We also found the tests to be unbiased in extensive simulation studies in the other settings of Tables I and II (data not shown). These results are also consistent with the fact that in all of the scenarios in Tables I and II where we had observed modest bias in parameter estimation, the direction of bias was always towards the null value of the parameters.

TABLE III. High-penetrant rare dominant gene: empirical type-I error rates of wald-tests with 5% significance levels

Methods	H_0: $\mathrm{MI}_{GE}=1$	H_0: $\mathrm{OR}(G\mid E=1)=1$	H_0: $\mathrm{OR}(E\mid G=1)=1$	H_0: $\mathrm{AI}_{GE}=0$
L_{CC}	0.042	0.048	0.040	0.036
L_{CCGP}	0.040	0.052	0.056	0.030

GENERAL FAMILY DATA WITH r AFFECTED AND s UNAFFECTED SUBJECTS

In this section, we briefly outline how the proposed methods for analyzing 1:1 family-matched case-control studies can be utilized for more general family studies that collect genotype and environmental exposure data for more than one affected and/or unaffected family members. Suppose there are M families sampled into a study and each family defines a matched set consisting of comparable cases and controls in the family, all of whom have data on both G and E. Suppose the i^{th} family defines a matched set consisting of r_i cases and s_i controls with a total of $n_i=r_i+s_i$ subjects. Let D_{i0} and D_{i1} denote the set of controls and cases, respectively, for the i^{th} family.

Liang [1987] proposed a pairwise pseudo-likelihood approach for analysis of matched case-control studies in which the contribution of a matched set of r_i cases and s_i controls is given by the product of the usual conditional likelihoods ($L_{i,\mathrm{CLR}}$) for 1:1 matched studies for all possible $r_i \times s_i$ case-control pairs within that matched set. Under the gene-environment independence assumption, we propose to use a similar pseudo-likelihood approach based on case-control pairs, except that for each pair we use an appropriate efficient conditional likelihood instead of the traditional conditional likelihood. More explicitly, the pseudo-likelihood for the data can be written in the general form

$$L=\prod_{i=1}^{M}\prod_{j\in D_{i0},k\in D_{i1}} L_{(jk)_i,*}, \qquad (10)$$

where $L_{(jk)_i,*}$ denotes an appropriate conditional- or pseudo-conditional likelihood for the $(j,k)^{\text{th}}$ case-control pair within the i^{th} matched set: the likelihood should be chosen efficiently according to whether and what kind of parental genotype data are available for that pair. In particular, when parental genotype data (both parents) are available for both subjects in the pair, $L_{(jk)_i,*}$ can be defined to be the conditional likelihood $L_{(jk)_i,\mathrm{CCGP}}$

(formula 8). When parental genotype data (both parents) are available only for the case in the pair, $L_{(jk)_i,*}$ can be defined to be $L_{(jk)_i,\mathrm{CC-CPT}} = L_{(jk)_i,\mathrm{CC}} \times L_{(jk)_i,\mathrm{CPT}}$, a pseudo-likelihood that efficiently combines information from case-control and case-parents-trio data. In all other cases, we propose to use $L_{(jk)_i,\mathrm{CC}}$ (formula 5). Finally, we observe that if for some families only cases and their parents, but no matching controls, are available, the contribution of these families in the above pseudo-likelihood can be defined by $L_{i,\mathrm{CPT}}$, the usual conditional-likelihood for case-parents-trio data.

In the pseudo-likelihood (10), the contribution of a matched set is obtained by taking the product of the contributions of all different case-control pairs within the set pretending that the different pairs from the same family are independent. From the theory of estimating-equations [Godambe, 1991], it is well known that such pseudo-likelihood methods produce consistent estimates of regression parameters even if in truth there is correlation among paired units within the same family. For variance estimation, however, the correlation within a family needs to be accounted for. A sandwich covariance matrix estimator that can account for such correlation is given in the Appendix: the formula is (A.9). In addition, bootstrap sampling with matched-sets as the sampling units can be used to obtain covariance matrix estimates that can account for within-family correlation.

DISCUSSION

We have proposed a new paradigm of conditional likelihood for analysis of family-based case-control studies. This approach, with a rare disease approximation, leads to a variety of simple but highly efficient methods of estimating statistical interaction and other risk parameters of interest involving genetic and environmental exposures. These methods exploit within family *G-E* independence assumptions that are much less stringent than the *G-E* independence assumption that has been previously utilized for case-only and population-based case-control studies. To the best of our knowledge, the proposed method involving the likelihood $L_{i,\mathrm{CC}}$ represents the first successful effort for exploiting the within-family ''Type-I independence'' assumption in the context of family-based case-control studies. Moreover, the likelihood $L_{i,\mathrm{CC}}$ can be used for efficient analysis of any other type of matched case-control study, the required assumption being that *G* and *E* are independent within ''matched subjects'' in the population. The proposed method involving the likelihood $L_{i,\mathrm{CCGP}}$, exploiting the ''Type-II independence'' assumption, represents the first approach to a unified analysis of data from family-based case-control studies that include parental genotype information.

An important aspect of the proposed general conditional likelihood ($L_{i,\mathrm{general}}$) is that it simultaneously conditions on a *set* phenotype event ($D_1 + D_0$) and a *set* genotype event (G). Historically, the idea of conditioning on a *set* phenotype event has been used in the setting of traditional conditional logistic regression ($L_{i,\mathrm{CLR}}$) analysis of matched case-control studies. In this approach, however, one conditions on individual covariate information of the case-control subjects. The idea of conditioning on a *set* genotype event has also existed for a while in the literature of family studies. The well known likelihood ($L_{i,\mathrm{CPT}}$) of case-parent-trio design is formed by conditioning on the parental genotypes ($\mathcal{G}^p$) of the cases, which yield the *set* of all possible genotypes for the offspring. Recent extensions of $L_{i,\mathrm{CPT}}$ for dealing with missing parental genotype information has also been based on conditioning on various types of *set* genotype events [Clayton, 1999; Cordell and Clayton 2002; Rabinowitz and Laird, 2000]. All of these methods, however, condition on individual phenotype information of the family members. In this article, we show how the two approaches of conditioning on a *set* genotype event and conditioning on a *set* phenotype event can be unified through the general conditional likelihood ($L_{i,\mathrm{general}}$), resulting in novel and efficient methods for analysis of matched case-control studies with or without parental genotype information.

Our simulation studies clearly demonstrate the efficiency advantage of our methods ($L_{i,\mathrm{CC}}$ and $L_{i,\mathrm{CCGP}}$) over the traditional conditional logistic regression method for analysis of family-based case-control and case-parents studies. These results also reveal some intriguing design implications. Several previous studies have compared the relative efficiencies of sibling-case-control (SCC) and case-parent trio (CPT) designs for estimation of the multiplicative interaction parameter: they generally concluded that while the former design tends to be superior for dominant genes, the latter design is more efficient for recessive genes [Witte et al., 1999; Gauderman, 2002]. However, in these studies the method employed for analysis of the CPT design implicitly assumes *G-E* independence,

but that for the SCC design does not exploit any such assumption. In our study, when we analyzed both designs using similar independence assumptions, we found not only that the efficiency advantage of the SCC design over CPT design for dominant genes is even greater than reported before, but also that the SCC design can be more efficient than the CPT design even for recessive genes. In terms of other parameters of interest, a weakness of the CPT design is that it cannot be used to estimate either the additive interaction or the association parameter for the environmental exposures. This design, however, is quite efficient for estimation of the genetic association parameter. The SCC design, on the other hand, although it produces an estimate of all different parameters of interest, was inefficient for estimation of the genetic association parameter $OR(G|E=1)$.

Our simulation studies also demonstrate the optimality of the family-based case-control design that includes parental genotype information. Although the potential promise of such a hybrid design has been discussed previously [Weinberg and Umbach, 2000], the actual efficiency of such a design has not been evaluated. It is worth noting that we compare efficiencies of different family-based designs with a fixed number of families, but different designs require different amounts of data collection within a family. The sibling-case-control design with parental genotype information, for example, requires one additional genotyping compared to an ordinary sibling-case-control design: note that the sibling control need not be genotyped if parental genotype data are available. It also requires one additional exposure assessment compared to an ordinary case-parents design. Given that family members of cases are usually well motivated, the effort required for such additional data collection may be worthwhile, considering the potential for large efficiency gain.

Comparison of the efficiency of the sibling-case-control designs (SCC and SCCGP) with that of the population-based case-control (PCC) design shows that our methodology increases the utility of the former design for a wider set of situations. Previous studies, as well as our simulations, demonstrate that the SCC design, when analyzed with traditional methods, generally tends to be less efficient than the PCC design, except when the genetic variant under study is rare and/or sibling correlation in the environment exposure is modest. When analyzed with our methods, the SCC design retains the efficiency advantage over the PCC design for estimation of interaction and some of the other parameters of interest in a wider range of values for the gene-frequency and the sibling-correlation parameters. The SCCGP design retained the efficiency advantage for an even wider range of parameter values. Of course, one can also increase efficiency of the PCC design, based on methods that can exploit G-E independence. The required population level independence assumption, however, is much stronger and more likely to be violated due to spurious association as described below.

Although exploiting *G-E* independence leads to major efficiency gains, some caution is needed for use of this assumption. The case-only estimate of interaction, which relies on a population-based *G-E* independence assumption, has been shown to be severely biased when the assumption is violated [Albert et al., 2001]. Even if no direct association exists, association between *G* and *E* in the population may arise, for example, due to hidden population sub-structure across which genotype and exposure frequency may vary or by influence of family history on an individual's behavior regarding established risk factors such as smoking.

Although the family-based independence assumptions we exploit are much less stringent in general, it is important to realize that the Type-I independence assumption we exploit for case-control studies is robust to spurious association only for the sibling-case-control design. If cousins are selected as controls, they may partially but not completely share ethnic and family history background with the cases; thus, the possibility of spurious association remains. One possible remedy for minimizing such bias is to consider a conditional *G-E* independence model that can adjust for co-factors *S*, such as the ethnic origins of the unrelated parents, by specifying the difference in genotype-frequencies between a pair of relatives as a parametric function of the differences in *S* between the relatives. We have found that under this relaxed independence assumption, the general conditional likelihood in formula (3) can be used with $\mathcal{G}=\mathcal{G}^S$, the unordered genotype information in a matched pair, to jointly estimate the regression parameters of interest and the additional parameters of the conditional independence model.

The *G-E* independence assumption can be violated also due to direct association between *G* and *E*. Genetic polymorphisms in the smoking

metabolism pathway, for example, may not only modify a subject's risk from smoking, but also can influence his/her level of addiction to smoking. When the plausibility of such direct association exists, the advantage of the case-control design is that it has the option of being analyzed by the standard conditional logistic regression method that does not require the independence assumption. The case-parents design, however, intrinsically relies on the independence assumption and thus can lead to biased parameter estimates.

Our novel conditional likelihood framework opens several areas of further research. We have assumed rare disease for the simplification of conditional likelihood calculations. From the derivation shown in the Appendix it can be seen that the precise assumption required here is that the probability of at least one disease occurrence in a pair of individuals is small for all combinations of risk factors, a slightly stronger assumption than that is typically made for an individual subject in population-based case-control studies. In our simulation study, where overall disease prevalence was 1% in the population, we observed no bias in testing and only modest bias in parameter estimation even in situations where the disease-risk was high for certain combinations of G and E. Future work, however, is needed to study the impact of the rare disease assumption for more common diseases. The general conditional likelihood of the form (3) is valid whether the disease is rare or not: for common diseases, however, these likelihoods would involve the family-specific intercept parameters (α_F). Thus, one way of relaxing the rare disease assumption would be to assume a parametric random effect model for the distribution of α_F across families and estimate the corresponding parameters from the conditional likelihoods themselves [Pfeiffer et al., 2002].

Similar to the traditional conditional-logistic-regression (CLR) analysis, the proposed conditional likelihood procedures assume disease status is conditionally independent within families. However, if the locus under study is in linkage disequilibrium with another disease susceptibility locus, such a conditional independence model may not capture the correlation within a family that contributes more than two members in the study [Weinberg and Umbach, 2000]. In such a situation, the proposed pair-wise pseudo-likelihood (10) together with the robust variance estimator (A.9) still remains a valid method for analysis of the data. Alternative methods for dealing with residual correlation that have been previously developed in the context of traditional CLR [Siegmund et al., 2000; Rieger et al., 2001], also could be adopted in the setting of the novel conditional likelihoods.

When studying the effect of a gene through haplotypes is of interest, the proposed methodologies can be applied for studying haplotype-environment interaction, but some extensions are required to deal with phase ambiguity. We have demonstrated how to exploit parental genotype information on case-control subjects when both parents of a subject are available to be recruited. In practice, however, genotype information may be available only for one parent for some case-control subjects. One way of efficiently utilizing incomplete parental genotype information would be to choose the conditioning event $\mathcal{G}$ in $L_{i,general}$ (formula 3) to be the minimal sufficient statistics for the gene-frequency parameters [Rabinowitz and Laird, 2000]. Various alternative strategies, such as modelling mating-type parameters [Kraft et al., 2004], that have been proposed in the past for dealing with missing parental genotypes in case-parents studies, could be also useful for developing extensions of the proposed methodologies. These and other extensions of the methodologies will be studied in future publications.

ACKNOWLEDGEMENTS

Research was supported by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

REFERENCES

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. 2001. Limitations of the case-only design for identifying gene-environment interaction. Am J Epidemiol 154:687–693.

Chatterjee N, Carroll RJ. 2005. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika (in press).

Clayton D. 1999. A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. Am J Hum Genet 65:1170–1177.

Clayton D, McKeigue PM. 2001. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 358:1356–1360.

Cordell HJ, Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data:

application to *HLA* in type 1 diabetes. Am J Hum Genet 70: 124–141.
Cordell HJ, Barratt BJ, Clayton DG. 2004. Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. Genet Epidemiol 26:167–185.
Curtis D. 1997. Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–333.
Gauderman WJ. 2002. Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med 21:35–50.
Gauderman WJ, Witte JS, Thomas DC. 1999. Family-based association studies. J N Cancer Inst 26:31–37.
Godambe VP. 1991. Estimating functions. Oxford: Oxford University Press.
Hsu L, Zhao LP, Aragaki C. 2000. A note on a conditional-likelihood approach for family-based association studies of candidate genes. Human Hered 50:194–200.
Khoury MJ, Beaty TH, Cohen BH. 1993. Fundamentals of genetic epidemiology. Oxford: Oxford University Press.
Kraft P, Palmer C, Woodward J, Turunen J, Minassian S, Paunio T, Lonnqvist J, Peltonen L, Sinsheimer J. 2004. RHD maternal-fetal genotype incompatibility and schizophrenia: Extending the MFG test to include multiple siblings and birth order. Eur J Hum Genet 12:192–198.
Liang KY. 1987. Extended Mantel-Haenzel estimating procedure for multivariate logistic regression models. Biometrics 43:289–299.
Pfeiffer R, Gail MH, Pee D. 2002. Inference for covariates that accounts for ascertainment and random genetic effects in family studies. Biometrika 88:933–948.
Piegorsch WW, Weinberg CR, Taylor JA. 1994. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. Stat in Med 13:153–162.
Rabinowitz D, Laird N. 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Human Hered 50:211–223.
Rieger RH, Kaplan NL, Weinberg CR. 2001. Efficient use of siblings in testing for linkage and association. Genet Epidemiol 20:175–191.
Schaid DJ. 1999. Case-parents design for gene-environment interaction. Genet Epidemiol 16:261–273.
Self SG, Longton G, Kopecky KJ, Liang KY. 1991. On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 47:53–61.
Siegmund KD, Langholz B, Kraft P, Thomas DC. 2000. Testing linkage disequilibrium in sibships. Am J Hum Genet 67:244–288.
Spielman RS, Ewens JE. 1998. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458.
Thomas DC. 2000. Case-parents design for gene-environment interaction by Schaid. Genet Epidemiol 19:461–463.
Thompson WD. 1991. Effect modifications and limits of biological inference from epidemiologic data. J Clin Epidemiol 44:221–232.
Umbach DM, Weinberg CM. 1997. Designing and analyzing case-control studies to exploit independence of genotype and exposure. Stat Med 16:1731–1743.
Umbach DM,Weinberg CM. 2000. The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet 66:251–261.
Weinberg CR, Umbach DM. 2000. Choosing a retrospective design to assess joint genetic and environmental contribution to risk. Am J Epidemiol 152:197–203.
Witte JS, Gauderman WJ, Thomas DC. 1999. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epidemiol 149:693–705.

APPENDIX

DERIVATION OF PROPOSED CONDITIONAL LIKELIHOOD IN THE GENERAL CASE

Recall that (D_{i0}, G_{i0}, E_{i0}) and (D_{i1}, G_{i1}, E_{i1}) are the data for the control and case, respectively. In addition, $\mathcal{G}_i$ is the conditioning event, and $\mathcal{H}_{\mathcal{G}_i}$ is the set of ordered pairs (G_{i1}, G_{i0}) that are consistent with the information in $\mathcal{G}_i$. We also denote by F_i the i^{th} family. Note that for the i^{th} pair,

$$\begin{aligned}\mathcal{L}_{i,\text{general}} =& \text{pr}(D_{i1}=1, D_{i0}=0, G_{i1}, G_{i0} | D_{i1}+D_{i0}=1, \mathcal{G}_i, E_{i1}, E_{i0}, F_i)\\ =& \text{pr}(D_{i1}=1, D_{i0}=0 | D_{i1}+D_{i0}=1, G_{i1}, G_{i0}, G_i, E_{i1}, E_{i0}, F_i)\\ & \times \text{pr}(G_{i1}, G_{i0} | D_{i1}+D_{i0}=1, \mathcal{G}_i, E_{i1}, E_{i0}, F_i)\\ =& \text{pr}(D_{i1}=1, D_{i0}=0 | D_{i1}+D_{i0}=1, G_{i1}, G_{i0}, E_{i1}, E_{i0}, F_i)\\ & \times \text{pr}(G_{i1}, G_{i0} | D_{i1}+D_{i0}=1, \mathcal{G}_i, E_{i1}, E_{i0}, F_i) = \mathcal{L}_{i1} \times \mathcal{L}_{i2}.\end{aligned}$$

The term $\mathcal{L}_{i1}$ is given in equation (2). It is easily seen that

$$\mathcal{L}_{i2} = \frac{\text{pr}\{D_{i1}+D_{i0}=1 | G_{i1}, G_{i0}, E_{i1}, E_{i0}, F_i\}\text{pr}(G_{i1}, G_{i0} | \mathcal{G}_i, E_{i1}, E_{i0}, F_i)}{\sum_{(G'_{i1}, G'_{i0}) \in H_{G_i}} \text{pr}\{D_{i1}+D_{i0}=1 | G'_{i1}, G'_{i0}, E_{i1}, E_{i0}, F_i\}\text{pr}(G'_{i1}, G'_{i0} | \mathcal{G}_i, E_{i1}, E_{i0}, F_i)}.$$

By assumption, G and E are independent given $\mathcal{G}_i$ and F_i, and in addition the conditioning event removes the family effect, so that $\text{pr}(G_{i1}, G_{i0} | \mathcal{G}_i, E_{i1}, E_{i0}, F_i) = \text{pr}(G_{i1}, G_{i0} | \mathcal{G}_i)$. We thus have that

$$\mathcal{L}_{i2} = \frac{\text{pr}\{D_{i1}+D_{i0}=1 | G_{i1}, G_{i0}, E_{i1}, E_{i0}, F_i\}\text{pr}(G_{i1}, G_{i0} | \mathcal{G}_i)}{\sum_{(G'_{i1}, G'_{i0}) \in \mathcal{H}_{\mathcal{G}_i}} \text{pr}\{D_{i1}+D_{i0}=1 | G'_{i1}, G'_{i0}, E_{i1}, E_{i0}, F_i\}\text{pr}(G'_{i1}, G'_{i0} | \mathcal{G}_i)}. \quad \text{(A.1)}$$

In addition, using (1), we have that

$$\begin{aligned}
&\mathrm{pr}(D_{i1}+D_{i0}=1|G_{i1},G_{i0},E_{i1},E_{i0},F_i)\\
&=\mathrm{pr}(D_{i1}=1,D_{i0}=0|G_{i1},G_{i0},E_{i1},E_{i0},F_i)+\mathrm{pr}(D_{i1}=0,D_{i0}=1|G_{i1},G_{i0},E_{i1},E_{i0},F_i)\\
&=\frac{\exp(\alpha_{F_i})\times\left[\exp\{m(G_{i1},E_{i1};\beta)\}+\exp\{m(G_{i0},E_{i0};\beta)\}\right]}{\left[1+\exp\{\alpha_{F_i}+m(G_{i1},E_{i1};\beta)\}\right]\times\left[1+\exp\{\alpha_{F_i}+m(G_{i0},E_{i0};\beta)\}\right]}\\
&\approx\exp(\alpha_{F_i})\times\left[\exp\{m(G_{i1},E_{i1};\beta)\}+\exp\{m(G_{i0},E_{i0};\beta)\}\right],
\end{aligned}\tag{A.2}$$

where the approximation in the last step is based on the assumption of rare disease. Thus, combining (A.1) and (A.2), we see that

$$\mathcal{L}_{i2}=\frac{\left[\sum_{j=0}^{1}\exp\{m(G_{ij},E_{ij},\beta)\}\right]\mathrm{pr}(G_{i1},G_{i0}|\mathcal{G}_i)}{\sum_{(G'_{i1},G'_{i0})\in\mathcal{H}_{\mathcal{G}_i}}\left[\sum_{j=0}^{1}\exp\left\{m(G'_{ij},E_{ij},\beta)\right\}\right]\mathrm{pr}(G'_{i1},G'_{i0}|\mathcal{G}_i)}.\tag{A.3}$$

Combining (2) and (A.3), we find that

$$\mathcal{L}_{i,\mathrm{general}}=\frac{\exp\{m(G_{i1},E_{i1},\beta)\}\mathrm{pr}(G_{i1},G_{i0}|\mathcal{G}_i)}{\sum_{(G'_{i1},G'_{i0})\in\mathcal{H}_{\mathcal{G}_i}}\left[\sum_{j=0}^{1}\exp\left\{m(G'_{ij},E_{ij},\beta)\right\}\right]\mathrm{pr}(G'_{i1},G'_{i0}|\mathcal{G}_i)},\tag{A.4}$$

which is the natural generalization of (7).

For the standard matched family-based case-control design where we set $\mathcal{G}_i$ to be the *set* of genotypes for the i^{th} case-control pair, under the assumption that $\mathrm{pr}(G_{i1}=g_1,G_{i0}=g_0|\mathcal{G}_i)=\mathrm{pr}(G_{i1}=g_0,G_{i0}=g_1|\mathcal{G}_i)$, we have that $\mathrm{pr}(G_{i1},G_{i0}|\mathcal{G}_i)=1/2$, thus verifying (5). For the parental genotype case-control design, we have already verified (7) simply by setting $\mathcal{G}_i=\mathcal{G}_i^P$, the full parental genotype information for the i^{th} case-control pair.

STANDARD ERROR ESTIMATION

In this section, we show that the asymptotic covariance matrix of the estimate, $\widehat{\beta}$, that maximizes (A.4) can be estimated as follows:

$$\mathrm{cov}(\widehat{\beta})=\left[\sum_{i=1}^{M}\left\{\frac{R_i(\widehat{\beta})}{A_i(\widehat{\beta})}-\frac{C_i(\widehat{\beta})C_i^{\mathrm{T}}(\widehat{\beta})}{A_i^2(\widehat{\beta})}\right\}\right]^{-1},\tag{A.5}$$

where if $m_\beta(G,E,\beta)$ and $m_{\beta\beta}(G,E,\beta)$ are the vector and matrix of first and second partial derivatives of $m(G,E,\beta)$ with respect to β, then

$$\begin{aligned}
A_i(\beta)&=\sum_{(g'_1,g'_0)\in\mathcal{H}_{G_i}}\left[\sum_{j=0}^{1}\exp\left\{m(g'_j,E_{ij},\beta)\right\}\right]\mathrm{pr}(G_{i1}=g'_1,G_{i0}=g'_0|\mathcal{G}_i);\\
C_i(\beta)&=\sum_{(g'_1,g'_0)\in\mathcal{H}_{G_i}}\left[\sum_{j=0}^{1}m_\beta(g'_j,E_{ij},\beta)\exp\left\{m(g'_j,E_{ij},\beta)\right\}\right]\mathrm{pr}(G_{i1}=g'_1,G_{i0}=g'_0|\mathcal{G}_i);\\
R_i(\beta)&=\sum_{(g'_1,g'_0)\in\mathcal{H}_{G_i}}\left[\sum_{j=0}^{1}m_\beta(g'_j,E_{ij},\beta)m_\beta^{\mathrm{T}}(g'_j,E_{ij},\beta)\exp\left\{m(g'_j,E_{ij},\beta)\right\}\right]\\
&\quad\times\mathrm{pr}(G_{i1}=g'_1,G_{i0}=g'_0|\mathcal{G}_i).
\end{aligned}$$

To show (A.5), an alternative formulation of (A.4) is useful. No longer insisting that D_{i1} is the case, so that $(D_{i1},D_{i0})=(1,0)$ or $(0,1)$, equation (A.4) actually shows that

$$\begin{aligned}&\mathrm{pr}(D_{i1}=d_1,D_{i0}=d_0,G_{i1}=g_1,G_{i0}=g_0|D_{i1}+D_{i0}=d_1+d_0=1,\mathcal{G}_i,E_{i1},E_{i0},F_i)\\&=\frac{\exp\{d_1m(g_1,E_{i1},\beta)+d_0m(g_0,E_{i0},\beta1)\}\mathrm{pr}(G_{i1}=g_1,G_{i0}=g_0|\mathcal{G}_i)}{A_i(\beta)}.\end{aligned}\tag{A.6}$$

This means that the derivative with respect to β of the loglikelihood for the i^{th} observation is

$$\ell_i(\beta)=-C_i(\beta)/A_i(\beta)+D_{i1}m_\beta(G_{i1},E_{i1},\beta)+D_{i0}m_\beta(G_{i0},E_{i0},\beta).\tag{A.7}$$

It is easily seen that the Hessian is

$$\begin{aligned}\ell_{i,\beta}(\beta)=&-\frac{R_i(\beta)}{A_i(\beta)}+\frac{C_i(\beta)C_i^{\mathrm{T}}(\beta)}{A_i^2(\beta)}+D_{i1}m_{\beta\beta}(G_{i1},E_{i1},\beta)+D_{i0}m_{\beta\beta}(G_{i0},E_{i0},\beta)\\&-\frac{\sum_{(g_1',g_0')\in\mathcal{H}_{\mathcal{G}_i}}\left[\sum_{j=0}^{1}m_{\beta\beta}(g_j',E_{ij},\beta)\exp\left\{m(g_j',E_{ij},\beta)\right\}\right]\mathrm{pr}(G_{i1}=g_1',G_{i0}=g_0'|\mathcal{G}_i)}{A_i(\beta)}.\end{aligned}$$

Using (A.6), it is easy to see that the expectation of the sum of the last three terms conditional on $(D_{i1}+D_{i0}=1,\mathcal{G}_i,E_{i1},E_{i0},F_i)$ equals zero, and hence that the expected Fisher information for the i^{th} observation is $R_i(\beta)/A_i(\beta)-C_i(\beta)C_i^{\mathrm{T}}(\beta)/A_i^2(\beta)$, thus proving (A.5).

ASYMPTOTIC EFFICIENCY THEORY

Both $\mathcal{L}_{i1}$ and $\mathcal{L}_{i2}$ are conditional distributions depending on the parameter β. Hence the derivatives of their logarithms $\mathcal{L}_{i1,\beta}(\beta)$ and $\mathcal{L}_{i2,\beta}(\beta)$ behave as if they are log-likelihood scores, and when summed over the data each have information matrices $I_1(\beta)$ and $I_2(\beta)$, respectively. The claim that our method is asymptotically more efficient than conditional logistic regression then follows if we can show that $E\{\mathcal{L}_{i1,\beta}(\beta)\mathcal{L}_{i2,\beta}^{\mathrm{T}}(\beta)\}=0$. To see this, note that by their definitions as derivatives of logarithms of conditional probabilities, and making the rare disease assumption, we have that

$$0=E\{\mathcal{L}_{i1,\beta}(\beta)|D_{i1}+D_{i0}=1,G_{i1},G_{i0},\mathcal{G}_i,E_{i1},E_{i0},F_i\}.\tag{A.8}$$

Of course,

$$E\left\{\mathcal{L}_{i1,\beta}(\beta)\mathcal{L}_{i2,\beta}^{\mathrm{T}}(\beta)\right\}=E\left[E\left\{\mathcal{L}_{i1,\beta}(\beta)\mathcal{L}_{i2,\beta}^{\mathrm{T}}(\beta)|D_{i1}+D_{i0}=1,G_{i1},G_{i0},\mathcal{G}_i,E_{i1},E_{i0},F_i\}\right],$$

where the interior expectation is with respect to the distribution of (D_{i1},D_{i0}) given $D_{i1}+D_{i0}=1$. However, $\mathcal{L}_{i2,\beta}(\beta)$ is a function only of $(D_{i1}+D_{i0}=1,G_{i1},G_{i0},G_i,E_{i1},E_{i0},F_i)$, and does not depend otherwise on (D_{i1},D_{i0}), so that by (A.8),

$$E\left\{\mathcal{L}_{i1,\beta}(\beta)\mathcal{L}_{i2,\beta}^{\mathrm{T}}(\beta)\right\}=E\left[E\{\mathcal{L}_{i1,\beta}(\beta)|D_{i1}+D_{i0}=1,G_{i1},G_{i0},\mathcal{G}_i,E_{i1},E_{i0},F_i\}\mathcal{L}_{i2,\beta}^{\mathrm{T}}(\beta)\right]=0.$$

This verifies the claim.

SANDWICH VARIANCE ESTIMATOR FOR PAIRWISE PSEUDO-LIKELIHOOD

Let $l_{(jk)_i}(\beta)=\partial\log\mathcal{L}_{(jk)_i,*}/\partial\beta$ and $V_{(jk)_i}(\beta)=\partial\log\mathcal{L}_{(jk)_i,*}/\partial\beta\partial\beta^T$ denote the vector of first-derivatives and matrix of second-derivatives, respectively, of the log-pseudo-likelihood for the $(jk)^{\text{th}}$ pair of the i^{th} family. The covariance matrix of $\widehat{\beta}$ then can be estimated by the sandwich estimator defined as follows:

$$\begin{aligned}\mathrm{cov}|(\widehat{\beta})=&Q_3^{-1}(\widehat{\beta})Q_4(\widehat{\beta})Q_3^{-1}(\widehat{\beta});Q_3(\beta)=\sum_{i=1}^{M}\sum_{j\in_{D_{i0}},k\in_{D_{i1}}}V_{(jk)_i}(\beta);\\Q_4(\beta)=&\sum_{i=1}^{M}\left\{\sum_{j\in_{D_{i0}},k\in_{D_{i1}}}\ell_{(jk)_i}(\beta)\right\}\left\{\sum_{j\in_{D_{i0}},k\in_{D_{i1}}}\ell_{(jk)_i}(\beta)\right\}^{\mathrm{T}};\end{aligned}$$